November 27, 1989 LIDS-P-1927

Zero Error Decision Feedback Capacity of Discrete Memoryless Channels*

İ. Emre Telatar† Robert G. Gallager‡

Abstract

The zero error decision feedback capacity for discrete memoryless channels is found using a constant composition coding argument. Surprisingly, the result is larger for many channels than the known bounds derived earlier by random coding arguments. An upper bound for the retransmission probability is also derived and is hoped to be asymptotically tight for the high rate region. For erasures and error schemes, we show that the bounds in [For68] are not tight and we improve these bounds to obtain new exponential error bounds for such schemes.

*This research was supported by the Army Research Office under grant DAAL03-86-K-0171 (Center for Intelligent Control Systems)

†Rm. 35-307, M.I.T. Laboratory for Information and Decision Systems, Cambridge MA 02139

‡Rm. 35-208, M.I.T. Laboratory for Information and Decision Systems, Cambridge MA 02139

1 Introduction

We will discuss communication systems in which a noiseless feedback link is available from the decoder to the encoder in addition to a discrete and memoryless forward link (Figure 1). Notice that the feedback link is from the output of the decoder, this type of feedback is

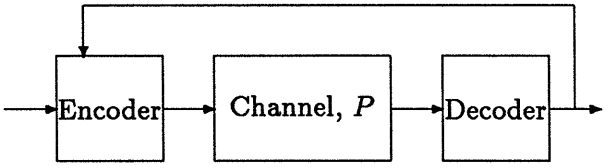

Figure 1: Decision Feedback Communication Model

called "decision feedback," and is weaker than that in which the transmitter is aware of the channel outputs. In this kind of systems, the decoder is given the option of not decoding (erasing), and requesting a retransmission. It is clear that if we use a block code of rate R over the channel and the erasure probability is say ϵ, then we can achieve an effective rate of $(1-\epsilon)R$. Thus, if the erasure probability can be made small, one can transmit without any loss of rate, but with potentially significant improvement on the error probability. Clearly, allowing more erasures will decrease the number of errors, what one is interested in this case is the tradeoff between the probabilities of error and erasure events. The first event occurs when the decoder chooses an incorrect message, and the other is when the decoder rejects all messages and declares an erasure. The optimal decoding rule and upper bounds on the error and erasure probabilities were found in [For68]; Csiszár and Korner [CK81] rederived these bounds using a constant composition coding technique. Forney in [For68] also notes that for a particular class of channels the erasure probability can be made arbitrarily small while maintaining *zero* error probability. For those channels there is a nonzero "zero error decision feedback capacity." A necessary and sufficient condition for this type of behavior is the existence of an output which can be reached by some but not all inputs. In [For68], Forney derives the following lower bound to this capacity through a random coding argument.

$$R_{0df}(P) = \max_{Q} -\sum_{j} w_j \ln v_j,$$

where Q runs over the space of single letter input probability distributions, j runs over the channel output alphabet, w_j denotes the probability of output j and v_j denotes the collective probability of inputs which lead to output j with positive probability.

In this paper we take a second look at zero error decision feedback capacity, and, using constant composition arguments, derive the actual capacity of the channel. Our approach will be the following:

- We will choose our codewords such that each has the same composition, i.e., each codeword is a permutation of every other codeword.

- A given output sequence will be decoded only if it can be generated from one and only one of the codewords. Otherwise an erasure will be declared and a repeat request will be issued.
- We will bound the probability of erasure by carving up the set of possible outputs from a given codeword into constant composition noise components.
- The largest rate for which the above bound can be made arbitrarily small will be a lower bound to the zero error decision feedback capacity. We will prove a converse to show that this lower bound is also an upper bound.

The capacity expression one comes up with after such a derivation is:

$$C_{0df}(P) = \max_{Q} \min_{P'} I(Q, P'),$$

where $I(Q, P')$ is the mutual information between the input and output ensembles when the input distribution is Q and the transition probabilities are P'. P' in the minimization ranges over all transition probability matrices which impose the same output distribution as P and also introduce no extra connections from inputs to outputs (i.e., if $P(j \mid k) = 0$ then so is $P'(j \mid k)$.) The maximization over Q ranges over all single letter input distributions.

1.1 Example

To illustrate the ideas above, and make the rest of the paper more transparent, let us analyze a simple example: The channel we will consider will be the Z-channel (see Figure 2), the transition probabilities of which are given by

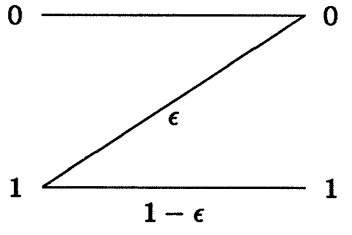

Figure 2: Z-Channel

$$\begin{aligned} P(0 \mid 0) &= 1, \qquad P(1 \mid 0) = 0, \\ P(0 \mid 1) &= \epsilon, \qquad P(1 \mid 1) = 1 - \epsilon. \end{aligned}$$

Let $C(Z_\epsilon)$ denote the (ordinary) capacity of this channel, and let $\mathbf{x}_1, \ldots, \mathbf{x}_M$ be the codewords of a block code of length n, rate $R = (\ln M)/n$ and maximum error probability λ. Let us classify the codewords according to their "type," in this case according to the number of 1's they have, which is their weight. Note that there can be at most $n + 1$ classes (each codeword has 0, 1,..., or n 1's), so the largest class contains at least $M/(n + 1)$

codewords. Consider eliminating all other codewords. The new code such formed has an error probability of at most λ, and rate at least

$$\frac{1}{n}\ln\frac{M}{n+1} = R - \frac{\ln(n+1)}{n}.$$

Thus, for large enough n, we can achieve any rate less than $C(Z_\epsilon)$ by a constant composition code, i.e., a code for which all codewords have the same number of 0's and 1's.

Now consider a constant composition code of rate R and a maximum likelihood decoder, with maximum probability of error λ. For a received sequence $\mathbf{y}$, define $N_{k,j}(\mathbf{x}_m, \mathbf{y})$ to be the number of i's such that $y_i = j$ and $x_{mi} = k$. Notice that $\mathbf{x}_m$ can lead to $\mathbf{y}$ with positive probability if and only if $N_{0,1}(\mathbf{x}_m, \mathbf{y}) = 0$. If this is the case $N_{1,1}(\mathbf{x}_m, \mathbf{y})$ is just the weight of $\mathbf{y}$ and $N_{1,0}(\mathbf{x}_m, \mathbf{y})$ is given by

$$(\#\text{ of 1's in } \mathbf{x}_m) - N_{1,1}(\mathbf{x}_m, \mathbf{y}) = (\text{weight of } \mathbf{x}_m) - (\text{weight of } \mathbf{y}),$$

which is independent of m, since the code is of constant composition. Since

$$\Pr(\mathbf{y} \mid \mathbf{x}_m) = \prod_{i=1}^{n} P(y_i \mid x_{mi}) = \begin{cases} \epsilon^{N_{1,0}(\mathbf{x}_m,\mathbf{y})}(1-\epsilon)^{N_{1,1}(\mathbf{x}_m,\mathbf{y})} & \text{if } N_{0,1}(\mathbf{x}_m, \mathbf{y}) = 0, \\ 0 & \text{else,} \end{cases}$$

we see that $\Pr(\mathbf{y} \mid \mathbf{x}_m)$ is either zero, or does not depend on m. Thus, for any $\mathbf{y}$, there is either a single codeword which can lead to it, or there is a set of equally likely codewords that the decoder must choose from. Since in the latter case the decoder will commit to an error with probability at least $1/2$, we have

$$\frac{1}{2}\Pr(\text{there is more than one candidate}) \leq \Pr(\text{Error}) \leq \lambda.$$

From this we can conclude that a decoder which erases whenever there is a chance of making an error has erasure probability less than 2λ.

The above argument demonstrates that one can achieve an arbitrarily small erasure probability while making no decoding errors for all rates up to the (ordinary) capacity of our example channel. Moreover, the exponential rate of decay of the erasure probability with increasing block length is the same as that of error probability in an ordinary scheme. For this case there is no need to prove a converse, since it well known [Sha56] that feedback cannot increase the ordinary capacity of a discrete memoryless channel.

2 Notation and Background

Let us first make the concept of constant composition precise and derive some results that will be used later. In doing so we will mainly follow the notation of [CK81].

Let $\wp(\mathcal{A})$ denote the set of probability distributions on a set $\mathcal{A}$. That is, we will write $Q \in \wp(\mathcal{K})$ to mean Q is a probability distribution on $\mathcal{K}$. Similarly, we will denote the set of probability transition matrices from a set $\mathcal{A}$ to a set $\mathcal{B}$ by $\wp(\mathcal{B} \mid \mathcal{A})$. We will denote the probability of a symbol $k \in \mathcal{K}$ by Q_k. For $j \in \mathcal{J}$ and $k \in \mathcal{K}$, P_{jk} will stand for the conditional probability of j given k.

Definition. *Given a set* $K = \{0, \ldots, K-1\}$, *and* $Q \in \wp(K)$, *we say a sequence* $\mathbf{x} \in K^n$ *is* Q-typical *if*

$$\forall k \in K \quad \frac{N_k(\mathbf{x})}{n} = Q_k,$$

where $N_k(\mathbf{x})$ *is the number of occurrences of* k *in the sequence* $\mathbf{x}$. *Conversely,* Q *is called the* type *of such an* $\mathbf{x}$. *The set of all* Q*-typical sequences,*

$$\{\mathbf{x} \in K^n : \mathbf{x} \text{ is } Q\text{-typical}\},$$

is denoted by T_Q^n.

The following lemma states that there is not "too many" types. That is to say, there is only polynomially many of them.

Lemma 2.1 *The number of distinct types* Q *on* K^n *is bounded by* $(n+1)^K$.

Proof. This follows from the fact that each $N_k(\mathbf{x})$ can take $n+1$ different values (including zero) for each k. □

Some properties of the Q-typical set are summarized below:

- The Q^n probability of an $\mathbf{x} \in T_Q^n$ is

$$Q^n(\mathbf{x}) = \prod_{i=1}^{n} Q(x_i) = \prod_k Q_k^{nQ_k} = \exp -nH(Q),$$

  where $H(Q) = -\sum_k Q_k \ln Q_k$.

- The Q^n probability of T_Q^n is then $Q^n(T_Q^n) = |T_Q^n| \exp -nH(Q)$. Thus we see that $|T_Q^n| \leq \exp nH(Q)$. The inequality can be expressed as a combinatoric fact:

$$\frac{n!}{\prod_k (nQ_k)!} \leq \exp nH(Q).$$

Definition. *Given another set* $J = \{0, \ldots J-1\}$, *and* $P \in \wp(J \mid K)$, *the distribution induced on* J *by* Q *and* P *is denoted by* PQ, *i.e.,*

$$(PQ)_j = \sum_k P_{jk} Q_k.$$

Given $\mathbf{x} \in T_Q^n$ *and* $P \in \wp(J \mid K)$, *we denote by* $T_P^n(\mathbf{x})$ *the set of all* $\mathbf{y} \in J^n$ *for which*

$$\frac{N_{k,j}(\mathbf{x},\mathbf{y})}{n} = Q_k P_{jk},$$

where $N_{k,j}(\mathbf{x},\mathbf{y})$ *is the number of occurrences of the pair* (k,j) *in the sequence* $\{(x_i, y_i)\}_{i=1}^n$. *The sequences in* $T_P^n(\mathbf{x})$ *are said to be* P-generated from $\mathbf{x}$.

The counting argument of Lemma 2.1 establishes that, for a given $\mathbf{x}$, the number of distinct sets $T_P^n(\mathbf{x})$ is bounded by $(n+1)^{KJ}$.

If $\mathbf{y}$ is in $T_P^n(\mathbf{x})$, then

$$P^n(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{n} P(y_i \mid x_i) = \prod_k \prod_j P_{jk}^{nQ_kP_{jk}} = \exp -nH(P \mid Q),$$

where $H(P \mid Q) = -\sum_k \sum_j Q_k P_{jk} \ln P_{jk}$. Similarly, for $\hat{P} \in \wp(J \mid K)$,

$$\hat{P}^n(\mathbf{y} \mid \mathbf{x}) = \exp -n\Big(H(P \mid Q) + D(P\|\hat{P} \mid Q)\Big),$$

where

$$D(P\|\hat{P} \mid Q) = \sum_k \sum_j Q_k P_{jk} \ln \frac{P_{jk}}{\hat{P}_{jk}}.$$

Consequently, we have

Lemma 2.2 $\hat{P}^n(T_P^n(\mathbf{x}) \mid \mathbf{x}) \leq \exp -nD(P\|\hat{P} \mid Q)$.

The following result from [CK81] states the existence of good codes with constant composition, that is, codes such that every codeword has the same type.

Lemma 2.3 *Given $R > 0$ and $Q \in \wp(K)$ such that $T_Q^n \neq \emptyset$ and $H(Q) > R$, then, for all $\delta > 0$, $\exists M \geq \exp n(R - \delta)$ and $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\} \subset T_Q^n$ such that $\forall P, \hat{P} \in \wp(J \mid K)$, and $\forall m$:*

$$\left| T_P^n(\mathbf{x}_m) \cap \Big(\bigcup_{m' \neq m} T_{\hat{P}}(\mathbf{x}_{m'}) \Big) \right| \leq |T_P^n(\mathbf{x}_m)| \exp -n|I(Q, \hat{P}) - R|^+,$$

provided $n \geq n_0(K, J, \delta)$, with $|x|^+ = \max\{0, x\}$ and

$$I(Q, \hat{P}) = \sum_k \sum_j Q_k \hat{P}_{jk} \ln \frac{\hat{P}_{jk}}{\sum_i Q_i \hat{P}_{ji}}.$$

Proof. See [CK81, pp. 162–164]. □

3 Zero Error Capacity with Decision Feedback

Given a channel with input alphabet K, output alphabet J and transition probabilities $P \in \wp(J \mid K)$, for each $\mathbf{x} \in K^n$, let $R_P^n(\mathbf{x})$ denote the set of all $\mathbf{y} \in J^n$ such that $P^n(\mathbf{y} \mid \mathbf{x}) > 0$. Notice that

$$R_P^n(\mathbf{x}) = \bigcup_{\hat{P}: P_{jk} = 0 \Rightarrow \hat{P}_{jk} = 0} T_{\hat{P}}^n(\mathbf{x}).$$

We will write the condition $P_{jk} = 0 \Rightarrow \hat{P}_{jk} = 0$ as $\hat{P} \ll P$. In measure theoretic terms, the condition states that $\hat{P}$ is absolutely continuous with respect to P. Given $Q \in \wp(K)$ with $T_Q^n \neq \emptyset$ and $H(Q) > R$, choose $\{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ as in Lemma 2.3. Then the following is the set of outputs $\mathbf{y}$ for which an erasure is declared when $\mathbf{x}_m$ is transmitted:

$$R_P^n(\mathbf{x}_m) \cap \Big(\bigcup_{m' \neq m} R_P^n(\mathbf{x}_{m'}) \Big) = \bigcup_{\substack{\hat{P}: \hat{P} \ll P \\ P': P' \ll P}} \Big[T_{\hat{P}}^n(\mathbf{x}_m) \cap \Big(\bigcup_{m' \neq m} T_{P'}^n(\mathbf{x}_{m'}) \Big) \Big].$$

Noticing that $T_{\hat{P}}^n(\mathbf{x}_m) \cap T_{P'}^n(\mathbf{x}'_m) = \emptyset$ whenever $P'Q \neq \hat{P}Q$, the union above can be restricted to those $\hat{P}, P'$ such that $P'Q = \hat{P}Q$. Using Lemmas 2.2 and 2.3, the probability of erasure conditional on $\mathbf{x}_m$ becomes

$$\begin{aligned} P^n\left\{ R_P^n(\mathbf{x}_m) \cap \left(\bigcup_{m' \neq m} R_P^n(\mathbf{x}_{m'}) \right) \,\middle|\, \mathbf{x}_m \right\} &\leq \sum_{\substack{\hat{P}:\hat{P} \ll P \\ P':P' \ll P,\, P'Q = \hat{P}Q}} e^{-n\left(D(\hat{P}\|P|Q) + |I(Q,P') - R|^+\right)} \\ &\leq (n+1)^{2KJ} \exp -nE(Q,P,R), \end{aligned}$$

where

$$E(Q,P,R) = \min_{\substack{\hat{P}:\hat{P} \ll P \\ P':P' \ll P,\, P'Q = \hat{P}Q}} \left[D(\hat{P}\|P \mid Q) + |I(Q,P') - R|^+ \right].$$

Notice that reliable communication is possible whenever the above minimum is positive. Setting this minimum equal to zero, we see that both terms involved in the expression above must be zero:

$$D(\hat{P}\|P \mid Q) = 0 \quad \text{and} \quad |I(Q,P') - R|^+ = 0. \tag{1}$$

Now, $D(\hat{P}\|P \mid Q) = 0$ if and only if $Q_k \hat{P}_{jk} = Q_k P_{jk}$ for all $k \in \mathcal{K}$, $j \in \mathcal{J}$. In particular, $D(\hat{P}\|P \mid Q) = 0$ implies $\hat{P}Q = PQ$. Recalling the constraints on P', we see that equation (1) will be satisfied if and only if

$$R \geq \min_{\substack{P':P' \ll P \\ P'Q = PQ}} I(Q,P'),$$

and the largest rate for which the exponent will remain positive is

$$C_{0df}(P) = \max_{Q} \min_{\substack{P':P' \ll P \\ P'Q = PQ}} I(Q,P').$$

We will show that C_{0df} as defined above is the zero error feedback capacity. In the above formula for C_{0df} notice that for any given Q, the feasible set

$$\left\{ P' \in \wp(\mathcal{J} \mid \mathcal{K}) \colon P' \ll P,\, P'Q = PQ \right\}$$

is a convex set and $I(Q,P')$ is convex $\cup$ in P'. However, the resulting function of Q to be maximized is not necessarily convex $\cap$ even though $I(Q,P')$ is.

3.1 Product Channels

In this section we will present results on channels that are built up from simpler channels via "products," as defined in [Sha56]. The result is especially important, as it throws some light on the nature of the optimization which yields the zero error decision feedback capacity, and it also eventually leads to the converse.

Given two channels with transition probabilities $\hat{P} \in \wp(\hat{\mathcal{J}} \mid \hat{\mathcal{K}})$ and $\tilde{P} \in \wp(\tilde{\mathcal{J}} \mid \tilde{\mathcal{K}})$, define the product of these two channels as a channel with the transition matrix $\hat{P} \times \tilde{P} \in \wp(\mathcal{J} \mid \mathcal{K})$, with

$$\mathcal{K} = \hat{\mathcal{K}} \times \tilde{\mathcal{K}}, \quad \mathcal{J} = \hat{\mathcal{J}} \times \tilde{\mathcal{J}},$$

$$(\hat{P} \times \tilde{P})_{(jl)(ki)} = \hat{P}_{jk}\tilde{P}_{li}.$$

That is, the product channel is the two channels used simultaneously.

We will need the following dual results for the proof of the third.

Lemma 3.1 *Let Q be a probability distribution on $\hat{\mathcal{K}} \times \tilde{\mathcal{K}}$ and let $\hat{Q}$ and $\tilde{Q}$ denote its marginal distributions on $\hat{\mathcal{K}}$ and $\tilde{\mathcal{K}}$. Then*

$$I(Q, \hat{P} \times \tilde{P}) \leq I(\hat{Q}, \hat{P}) + I(\tilde{Q}, \tilde{P}).$$

Proof. Since $I(Q, \hat{P} \times \tilde{P}) = H\Big((\hat{P} \times \tilde{P})Q\Big) - H(\hat{P} \times \tilde{P} \mid Q)$, and since

$$H(\hat{P} \times \tilde{P} \mid Q) = H(\hat{P} \mid \hat{Q}) + H(\tilde{P} \mid \tilde{Q}),$$

$$H\Big((\hat{P} \times \tilde{P})Q\Big) \leq H(\hat{P}\hat{Q}) + H(\tilde{P}\tilde{Q}),$$

the proof follows. □

Lemma 3.2 *Let $P = \hat{P} \times \tilde{P}$, $Q = \hat{Q} \times \tilde{Q}$. Then the minimization* $\min_{\substack{P' : P'Q = PQ \\ P' \ll P}} I(Q, P')$ *is achieved by some $P' = \hat{P}' \times \tilde{P}'$.*

Proof. Given P' with $P'Q = PQ$, $P' \ll P$, let

$$\hat{P}'_{jk} = \sum_{i \in \tilde{\mathcal{K}}, l \in \tilde{\mathcal{J}}} P'_{(jl)(ki)} \tilde{Q}_i,$$

$$\tilde{P}'_{li} = \sum_{k \in \hat{\mathcal{K}}, j \in \hat{\mathcal{J}}} P'_{(jl)(ki)} \hat{Q}_k.$$

It is easy to check that $\hat{P}'\hat{Q} = \hat{P}\hat{Q}$, $\tilde{P}'\tilde{Q} = \tilde{P}\tilde{Q}$, $\hat{P}' \ll \hat{P}$ and $\tilde{P}' \ll \tilde{P}$. Thus, $\hat{P}' \times \tilde{P}' \ll P$ and $(\hat{P}' \times \tilde{P}')Q = PQ$. Then

$$\begin{aligned}
I(Q, \hat{P}' \times \tilde{P}') - I(Q, P') &= H(P' \mid Q) - H(\hat{P}' \mid \hat{Q}) - H(\tilde{P}' \mid \tilde{Q}) \\
&= \sum_{k \in \hat{\mathcal{K}}} \sum_{i \in \tilde{\mathcal{K}}} \sum_{j \in \hat{\mathcal{J}}} \sum_{l \in \tilde{\mathcal{J}}} \hat{Q}_k \tilde{Q}_i P'_{(jl)(ki)} \ln \frac{\hat{P}'_{jk}\tilde{P}'_{li}}{P'_{(jl)(ki)}} \\
&\leq \sum_{k \in \hat{\mathcal{K}}} \sum_{i \in \tilde{\mathcal{K}}} \sum_{j \in \hat{\mathcal{J}}} \sum_{l \in \hat{\mathcal{J}}} \hat{Q}_k \tilde{Q}_i P'_{(jl)(ki)} \left[\frac{\hat{P}'_{jk}\tilde{P}'_{li}}{P'_{(jl)(ki)}} - 1 \right] \\
&= 0.
\end{aligned}$$

Hence the product distributions do better than any other distribution. □

Theorem 3.1 $C_{0df}(\hat{P} \times \tilde{P}) = C_{0df}(\hat{P}) + C_{0df}(\tilde{P})$.

Proof. By Lemma 3.1 we have

$$I(Q, \hat{P} \times \tilde{P}) \leq I(\hat{Q}, \hat{P}) + I(\tilde{Q}, \tilde{P}).$$

Then

$$\begin{aligned} C_{0df}(\hat{P} \times \tilde{P}) &= \max_{Q} \min_{\substack{P' : P' \ll \hat{P} \times \tilde{P} \\ P'Q = PQ}} I(Q, P') \\ &\leq \max_{Q} \min_{\substack{\hat{P}', \tilde{P}' : \hat{P}' \ll \hat{P}, \tilde{P}' \ll \tilde{P} \\ \hat{P}'\hat{Q} = \hat{P}\hat{Q}, \tilde{P}'\tilde{Q} = \tilde{P}\tilde{Q}}} I(Q, \hat{P}' \times \tilde{P}') \\ &\leq \max_{\hat{Q}, \tilde{Q}} \min_{\substack{\hat{P}', \tilde{P}' : \hat{P}' \ll \hat{P}, \tilde{P}' \ll \tilde{P} \\ \hat{P}'\hat{Q} = \hat{P}\hat{Q}, \tilde{P}'\tilde{Q} = \tilde{P}\tilde{Q}}} I(\hat{Q}, \hat{P}') + I(\tilde{Q}, \tilde{P}') \\ &= C_{0df}(\hat{P}) + C_{0df}(\tilde{P}), \end{aligned}$$

where the first step follows by restricting the range of the minimization. The converse inequality follows by restricting the maximization to Q's of the form $Q = \hat{Q} \times \tilde{Q}$ and applying Lemma 3.2 to see that

$$\begin{aligned} C_{0df}(\hat{P} \times \tilde{P}) &\geq \max_{Q = \hat{Q} \times \tilde{Q}} \min_{\substack{P' : P' \ll \hat{P} \times \tilde{P} \\ P'Q = (\hat{P} \times \tilde{P})Q}} I(Q, P') \\ &= \max_{\hat{Q}, \tilde{Q}} \min_{\substack{\hat{P}', \tilde{P}' : \hat{P}' \ll \hat{P}, \tilde{P}' \ll \tilde{P} \\ \hat{P}'\hat{Q} = \hat{P}\hat{Q}, \tilde{P}'\tilde{Q} = \tilde{P}\tilde{Q}}} I(\hat{Q}, \hat{P}') + I(\tilde{Q}, \tilde{P}') \\ &= C_{0df}(\hat{P}) + C_{0df}(\tilde{P}), \end{aligned}$$

thus completing the proof. □

3.2 The Converse

Assume that there exists a block code of length n with M codewords $\mathbf{x}_1, \ldots, \mathbf{x}_M$, which achieves zero error probability while erasure probability is less than ϵ. Consider the channel P^n and an input distribution

$$Q(\mathbf{x}) = \begin{cases} 1/M & \text{if } \mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}, \\ 0 & \text{else.} \end{cases}$$

Let us define the sets $D_m = R_P^n(\mathbf{x}_m) \setminus \left(\bigcup_{m' \neq m} R_P^n(\mathbf{x}_{m'}) \right)$, $m = 1, \ldots, M$. Notice that D_m is the set of outputs which can be reached only from $\mathbf{x}_m$. Also define $E = J^n \setminus \left(\bigcup D_m \right)$. Note that $P^n(D_m \mid \mathbf{x}_m) \geq 1 - \epsilon$ and for $m' \neq m$, $P^n(D_m \mid \mathbf{x}_{m'}) = 0$.

Let P' satisfy $P' \ll P^n$ and $P'Q = P^nQ$. Then for any $\mathbf{y} \in D_m$,

$$(P^nQ)(\mathbf{y}) = P^n(\mathbf{y} \mid \mathbf{x}_m)/M = P'(\mathbf{y} \mid \mathbf{x}_m)/M = (P'Q)(\mathbf{y}),$$

thus $P'(\mathbf{y} \mid \mathbf{x}_m) = P^n(\mathbf{y} \mid \mathbf{x}_m)$. Also if $m' \neq m$, $P' \ll P^n$ implies $P'(\mathbf{y} \mid \mathbf{x}_{m'}) = 0$. Thus for $m \neq m'$,

$$P'(D_m \mid \mathbf{x}_m) \geq 1 - \epsilon \quad \text{and} \quad P'(D_m \mid \mathbf{x}_{m'}) = 0.$$

Since partitioning the output space J^n into $D_1, \ldots, D_M$ and E will not increase the mutual information,

$$I(Q, P') \geq (1 - \epsilon) \ln M.$$

Since this is true for any P' with $P' \ll P^n$ and $P'Q = P^nQ$, we have

$$(1 - \epsilon) \ln M \leq \min_{\substack{P' : P' \ll P^n \\ P'Q = P^nQ}} I(Q, P') \leq C_{0fd}(P^n) = nC_{0fd}(P).$$

Thus $R = \ln M/n \leq (1 - \epsilon)^{-1} C_{0df}(P)$, establishing the converse.

3.3 Comparison with Single Letter Random Coding

As mentioned earlier, a lower bound to zero error decision feedback capacity of a discrete memoryless channel with transition probabilities $P \in \wp(J \mid K)$ using (single letter) random coding is given by $R_{0df}(P) = \max_{Q \in \wp(K)} Z(Q, P)$, where

$$Z(Q, P) = -\sum_j (PQ)_j \ln V(Q, P)_j,$$

with $V(Q, P)_j = \sum_{k : P_{jk} > 0} Q_k$. This expression gives the best possible performance of a system where each letter of each codeword is chosen independently with identical distribution.

Remark 3.1 *$Z(Q, P) \leq I(Q, P)$.*

Proof. Upon noting

$$\begin{aligned} Z(Q, P) - I(Q, P) &= \sum_j \sum_k Q_k P_{jk} \ln \frac{(PQ)_j}{V(Q, P)_j P_{jk}} \\ &= \sum_j \sum_{k : P_{jk} > 0} Q_k P_{jk} \ln \frac{(PQ)_j}{V(Q, P)_j P_{jk}} \\ &\leq \sum_j \sum_{k : P_{jk} > 0} Q_k P_{jk} \left[\frac{(PQ)_j}{V(Q, P)_j P_{jk}} - 1 \right] \\ &= \sum_j (PQ)_j - 1 = 0, \end{aligned}$$

we see that $Z(Q, P) \leq I(Q, P)$. □

Remark 3.2 *If Q, P and P' satisfy $PQ = P'Q$ and $P' \ll P$, then $Z(Q, P) \leq Z(Q, P')$.*

Proof. Since $P' \ll P$ implies $V(Q, P')_j \leq V(Q, P)_j$, we have

$$Z(Q, P) = -\sum_j (PQ)_j \ln V(Q, P)_j \leq -\sum_j (P'Q)_j \ln V(Q, P')_j = Z(Q, P').$$

□

Theorem 3.2 $R_{0df}(P) \leq C_{0df}(P)$.

Proof. By the above two remarks, for Q, P and P' such that $PQ = P'Q$ and $P' \ll P$,

$$Z(Q,P) \leq Z(Q,P') \leq I(Q,P').$$

Thus

$$Z(Q,P) \leq \min_{\substack{P':P' \ll P \\ P'Q=PQ}} I(Q,P'),$$

and hence

$$R_{0df}(P) = \max_Q Z(Q,P) \leq \max_Q \min_{\substack{P':P' \ll P \\ P'Q=PQ}} I(Q,P') = C_{0df}(P).$$

□

Therefore, the constant composition argument never leads to a capacity worse than that obtained via random coding. There are cases for which the inequality in the theorem is strict. A simple example is the Z-channel shown in Figure 2. For this channel, the set

$$\left\{P' : P' \ll P,\ P'Q = PQ\right\}$$

contains only P for nontrivial Q's. Thus, the decision feedback zero error capacity is equal to the ordinary capacity of the channel. The random coding argument however, gives an inferior answer of $(1-\epsilon)/e$.

A simple observation about $R_{0df}(P)$ is that

$$R_{0df}(\hat{P} \times \tilde{P}) \geq R_{0df}(\hat{P}) + R_{0df}(\tilde{P}),$$

rather than equality. So the natural thing to do would be to define

$$R_{0dfb}(P) = \lim_{n \to \infty} \frac{R_{0df}(P^n)}{n}$$

as the random coding zero error capacity. Theorems 3.1 and 3.2 show that

$$R_{0dfb}(P) \leq C_{0df}(P).$$

The next section will show that $R_{0dfb}(P)$ is in fact equal to $C_{0df}(P)$.

4 Generalization to Erasures and Errors Schemes

In the previous sections, we considered a decoder with a rather limited freedom. Namely, our decoder had to erase whenever it was not absolutely sure about the transmitted message. This is one extreme of the more general problem of erasures and errors decoding, our special case resulting from forcing the tradeoff between the erasures and errors in the direction of no errors. The other extreme of no erasures is the well studied classical communication problem.

In [For68] Forney studies erasures and errors decoding and the list decoding techniques, the latter allowing the decoder to produce more than one estimate. Forney shows that the optimal decoding rule is to decode m if

$$\frac{P^n(\mathbf{y} \mid \mathbf{x}_m)}{\sum_{m' \neq m} P^n(\mathbf{y} \mid \mathbf{x}_{m'})} \geq \exp(nT),$$

where T is an arbitrary parameter. If there is no m satisfying the above then the decoder declares an erasure. Note that for $T > 0$, at most one m can satisfy the decoding criterion, so this corresponds to erasures and errors decoding, and, for $T < 0$, there is at least one such m, corresponding to list decoding schemes.

Consider the above decoding rule with $T > 0$, so that we are considering erasures and errors decoding. If E_1 denotes the event of an erasure of error and E_2 denotes the event of an error, Forney shows for this decoding scheme

$$\Pr[E_1] \leq \exp -nE_1(R,T), \quad \text{and} \quad \Pr[E_2] \leq \exp -nE_2(R,T),$$

where

$$E_1(R,T) = \max_{Q, 0 \leq s \leq \rho \leq 1} [E_0(s,\rho,Q,P) - \rho R - sT] \tag{2}$$

$$E_2(R,T) = E_1(R,T) + T \tag{3}$$

and

$$E_0(s,\rho,Q,P) = -\ln F_0(s,\rho,Q,P)$$

$$F_0(s,\rho,Q,P) = \sum_j \left[\sum_k Q(k) P(j \mid k)^{1-s}\right] \left[\sum_{k'} Q(k') P(j \mid k')^{s/\rho}\right]^{\rho}.$$

With $T < 0$, $\exp -nE_2(R,T)$ is an upper bound to the average number of incorrect symbols on the list, and $\exp -nE_2(R,T)$ is an upper bound to the average number of incorrect symbols on the list.

We will improve on these bounds by applying this formula to the channel representing the n-fold use of the original one: The input alphabet is $\mathcal{K}^n$, whose elements we denote by $\mathbf{x} = (x_1, \ldots, x_n)$, the output alphabet is $\mathcal{J}^n$ with elements $\mathbf{y} = (y_1, \ldots, y_n)$, and the probability transitions are given by

$$P^n(\mathbf{y} \mid \mathbf{x}) = \prod_{i=1}^{n} P(y_i \mid x_i).$$

Let $Q \in \wp(\mathcal{K})$ and let $\mathbf{Q} \in \wp(\mathcal{K}^n)$ be the distribution

$$\mathbf{Q}(\mathbf{x}) = \begin{cases} 1/|T_Q^n| & \text{for } \mathbf{x} \in T_Q^n, \\ 0 & \text{else.} \end{cases}$$

Then $F_0(s,\rho,\mathbf{Q},P^n)$ becomes

$$F_0(s,\rho,\mathbf{Q},P^n) = \sum_{\mathbf{y} \in \mathcal{J}^n} \left[\sum_{\mathbf{x} \in T_Q^n} |T_Q^n|^{-1} P^n(\mathbf{y} \mid \mathbf{x})^{1-s}\right] \left[\sum_{\mathbf{x}' \in T_Q^n} |T_Q^n|^{-1} P^n(\mathbf{y} \mid \mathbf{x}')^{s/\rho}\right]^{\rho}.$$

Interchanging the sum over $\mathbf{y}$ and $\mathbf{x}$, partitioning the output space J^n into $\bigcup_{\hat{P}\in\mathcal{P}(J|K)} T_{\hat{P}}(\mathbf{x})$,

$$= |T_Q^n|^{-(1+\rho)} \sum_{\mathbf{x}\in T_Q^n} \sum_{\hat{P}\in\mathcal{P}(J|K)} \sum_{\mathbf{y}\in T_{\hat{P}}(\mathbf{x})} P^n(\mathbf{y}\mid\mathbf{x})^{1-s}\Big[\sum_{\mathbf{x}'\in T_Q^n} P^n(\mathbf{y}\mid\mathbf{x}')^{s/\rho}\Big]^{\rho},$$

and since

$$P^n(\mathbf{y}\mid\mathbf{x}) = \begin{cases} \exp -n(H(\hat{P}\mid Q) + D(\hat{P}\|P\mid Q)) & \mathbf{y}\in T_{\hat{P}}(\mathbf{x}), \hat{P}\ll P, \\ 0 & \mathbf{y}\in T_{\hat{P}}(\mathbf{x}), \hat{P}\not\ll P, \end{cases}$$

$$= |T_Q^n|^{-(1+\rho)} \sum_{\mathbf{x}\in T_Q^n} \sum_{\hat{P}\ll P} e^{-n(1-s)(H(\hat{P}|Q)+D(\hat{P}\|P|Q))} \sum_{\mathbf{y}\in T_{\hat{P}}^n(\mathbf{x})} \Big[\sum_{\mathbf{x}'\in T_Q^n} P^n(\mathbf{y}\mid\mathbf{x}')^{s/\rho}\Big]^{\rho}.$$

To evaluate the innermost summation for any given $\mathbf{y}\in T_{\hat{P}}^n(\mathbf{x})$, let us partition T_Q^n as follows:

$$T_Q^n = \bigcup_{P'} W(Q,P',\mathbf{y})$$

where $W(Q,P',\mathbf{y}) = \{\mathbf{x}'\in T_Q^n : \mathbf{y}\in T_{P'}^n(\mathbf{x}')\}$. Notice that the size of $W(Q,P',\mathbf{y})$ depends on $\mathbf{y}$ only through the type of $\mathbf{y}$, in our case $\hat{P}Q$, and if $P'Q\neq\hat{P}Q$ then this size is zero. Otherwise, $P'Q=\hat{P}Q$, and then

$$\begin{aligned} |W(Q,P',\mathbf{y})| &= \prod_j \frac{(n(\hat{P}Q)_j)!}{\prod_k (nQ_kP'_{jk})!} \\ &= \prod_j \frac{(n(P'Q)_j)!}{\prod_k (nQ_kP'_{jk})!} \\ &\le e^{n(H(Q)-I(Q,P'))}. \end{aligned}$$

Also note that if $P'\not\ll P$, $P(\mathbf{y}\mid\mathbf{x}')=0$ for all $\mathbf{x}'\in W(Q,P',\mathbf{y})$.

With these considerations, and noting that $|T_{\hat{P}}^n(\mathbf{x})|\le e^{nH(\hat{P}|Q)}$, we get

$$\begin{aligned} F_0(s,\rho,Q,P) \le\ & (n+1)^{2KJ} \exp -n \min_{\substack{\hat{P}\ll P, P'\ll P \\ P'Q=\hat{P}Q}} \Big\{ D(\hat{P}\|P\mid Q) + \rho I(Q,P') \\ & + s\Big(H(P'\mid Q) - H(\hat{P}\mid Q) + D(P'\|P\mid Q) - D(\hat{P}\|P\mid Q)\Big)\Big\} \end{aligned}$$

and thus,

$$\begin{aligned} E_{0b}(s,\rho,Q,P) &= \lim_{n\to\infty} \frac{-\ln F_0(s,\rho,Q,P^n)}{n} \\ &\ge \min_{\substack{\hat{P}\ll P, P'\ll P \\ P'Q=\hat{P}Q}} \Big\{ s\Big(H(P'\mid Q) - H(\hat{P}\mid Q) + D(P'\|P\mid Q) - D(\hat{P}\|P\mid Q)\Big) \\ &\quad + \rho I(Q,P') + D(\hat{P}\|P\mid Q)\Big\}. \end{aligned}$$

On examining the bounding technique, we see that all the bounds we have used are exponentially tight, and thus the inequality above is in fact an equality. Now one can use E_{0b}

instead of E_0 in Forney's formulas (2) and (3), and get better upper bounds on $\Pr[E_1]$ and $\Pr[E_2]$.

Note that the parameter T governs the tradeoff between erasures and errors. Letting $T \rightarrow \infty$ corresponds to the case of no errors. In this case the optimal s in calculating $E_1(R)$ and $E_2(R)$ is zero, and one can see that the highest rate for which $E_1(R)$ remains positive is $C_{0df}(P)$.

References

[CK81] Imre Csiszár and Janos Korner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 1981.

[For68] G. David Forney. Exponential error bounds for erasure, list and decision feedback schemes. *IEEE Transactions on Information Theory*, IT-14:206–220, March 1968.

[Sha56] Claude E. Shannon. The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, IT-2:8–19, September 1956.